DOCUMENT RESUME

ED 193 310 TM 800 625

AUTHOR O'Triel, Frances S.; Terry, B. Diane
TITLE A New Instrument for Measuring Teacher Effectiveness.
PUB DATE Mar 78
NOTE 14p.; Paper presented at the Annual Meeting of the Southeastern Psychological Association (Atlanta, GA, March, 1978).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluation Criteria; *Expectation; Higher Education; *Q Methodology; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Test Construction; Test Validity

ABSTRACT

A 24-item Q-sort faculty evaluation instrument based on Spady's model of teacher competency was constructed and piloted. Following the second pilot and factor analysis of the data, the items were reduced to 16. Test-retest reliability was .81. High correlations between items and their subscale total and low interitem correlations indicate independence of items. The instrument is administered at the beginning of the semester according to students' perceptions of an ideal professor for the course and at the end as a measure of the students' perception of the professor of the course. Comparison of pre-post results indicates professor's performance in relation to students' expectancies. Since these are known at the beginning of the course, the instructor may modify his/her approach early in the semester or explain to students why they are going to get something different from their anticipations. Evidence in the literature of the importance of expectancy in evaluation and results of this investigation suggest that current instruments may be invalid as measures of instructor performance because the variable is expectancy and is not controlled for. (Author/GSK)

A New Instrument for Measuring Teacher Effectiveness

Frances S. O'Tuel
University of South Carolina

B. Diane Terry
University of South Carolina

Paper presented at the Southeastern Psychological Association, Atlanta, Georgia March, 1978

A New Instrument for Measuring Teacher Effectiveness

Frances S. O'Tuel and B. Diane Terry
University of South Carolina

Student evaluations of instructors and/or instruction have been collected and used in various ways since the late 1800's. They seldom were allowed to influence critical decisions such as promotion, salary or retention. This has changed in the past two decades partially because of the thrust of the accountability movement and partially because of the desire of those making administrative decisions to have some "evidence" to support their hiring, firing, promoting and tenuring policies. Out of this, disputes have arisen as to the adequacy of the measurement properties of these student ratings and the use of student evaluations has already found its way to the courts (Herron, 1976).

Criticisms of student evaluation forms have centered around the reliability and validity issues of measurement and the absence of a theoretical framework.

Reliability is frequently numerically modest when computed for affective instruments. Student ratings are no exception; in most cases no test-retest reliabilities are reported and internal consistency is frequently absent. If student opinion is unreliable or if no one has tested to discover whether it is consistent, such results should not be used for decision making and probably should not even be used by instructors for their own information.

Validity questions are as unanswered as reliability questions. What do student evaluations measure is as yet unanswered. There is a

lack of agreement as to what would establish validity. The relation between ratings and student achievement is an obvious possibility; the studies are equivocal. Cohen and Berger (1970), Elliott (1949), Centra (1977), Frey (1973), and others have reported a positive relationship between student ratings and student achievement. However, a recent study by Measurement Services Center (1977) showed no relationship between the two and odin and Rodin (1972) found an inverse relationship. Ratings have been correlated with students' election to take more course work in the content field (McKeachie and Solomon, 1958), positive; with the age, experience, tenure, authorship of instructors (McDaniel and Feldhusen, 1970; MSC, 1977), not significant or negative; with size of class (Perry and Baumann, 1973), negative correlation; with elective or required course status (Gage, 1961), positive for elective; with level of course (Perry and Baumann, 1973; Gage, 1961), higher level course, higher ratings; with grade-giving history and nature of course (Bassin, 1974), lower grades given lower ratings and quantitative courses rated lower; with students' expected grade (Bausell and Magoon, 1972; Blum, 1936; Garverick and Carter, 1962; Parrish, et al, 1977), equivocal; with student GPA (Parrish, et al, 1977), negative; with student satisfaction with grading procedures (Shingles, 1977), positive; with number of times a week the class meets and time of day (Parrish, et al, 1977), negative with times met, more positive for class meetings later in day. Also studies have looked at whether ratings across subjects and different teaching methods should be compared. There appears to be some generalized agreement among teachers, students, and colleagues as to characteristics which should be evaluated so that with some caution this may be legitimate

(MSC, 1977; Treffinger and Feldhusen, 1970). Perhaps, the most serious validation question arises from this latter inquiry. If a generalized precourse rating of other courses at an institution predicts student evaluations of instructors, are we measuring instructor effectiveness at all? A similar question was posed in Ware and William's (1976) further exploration of the Dr. Fox effect which suggested that student ratings are reflecting instructor enthusiasm rather than any other qualities of the instructional process.

Factoral studies have attempted to establish a base for measurement and have supported the multidimensional characteristics of the ratings. From two to six factors have been found (Gibb, 1955; Isaacson, et al, 1964; Berger and Cohen, 1969; McKeachie, 1969; Meredith, 1969; Finkbeiner, et al, 1973; Jaeger and Friejo, 1974; Mazer, 1977; Shingles, 1977. Construction of an evaluation instrument which has an a priori theoretical framework seems absent in the literature. Halo effects, response set and socially desirable responses have clouded results. Winne (1977) explains inconsistencies in the findings of nominally similar treatments as aptitude treatment interactions (ATI). He states, "for example, students' preferences for one or another kind of teaching may influence learning and attitudes differently when the teaching they receive corresponds to their preferences versus when it does not (p. 390)."

A student evaluation instrument with a theoretical base and reliability and validity potential was constructed. Studies such as Treffinger and Feldhusen (1970), Winne (1977) and Shingles (1977) support the hypothesis that there are confounding variables in the

student characteristics which need to be controlled or measured; the Student Expectancy Evaluation (SEE) addresses these issues.

Theoretical Framework

Spady (1974) postulates four areas of competency in teachers. He states that teachers need to have something of substance and interest to say, which has been labeled subject matter expertise; to be capable of saying it clearly and accurately, which is a componetn of pedagogical expertise; to be able to say it in a stimulating and exciting fashion, which he refers to as charisma; and to base this communication directly on a concern for the personal welfare of each student, which is labeled empathy. Spady proposed that the most important component of a teacher's repertory of abilities is the capacity to establish a sense of rapport with students by caring about them as individuals. His position was that the critical variable was empathy and concern because as students mature (level of development and expectation), teacher's charisma and areas of expertise will erode in value. Spady's four areas of effectiveness - subject matter expertise, pedagogical expertise, charisma, and empathy - are the hypothetical constructs for the SEE. As to which was the most critical had not been empirically established. Tentatively in contrast to Spady's view, it would seem that as students advanced in their studies subject matter expertise would increase in importance and empathy would lose some of its inpact. Data from such a study are currently being analyzed.

Instrument Construction

From more than 30 evaluation instruments, 347 items were placed in an item pool. These were reduced to 76 positive statements by eliminating items which were similar to one another (just noticeable difference). The items were placed in four categories corresponding to Spady's (1974) four dimensions. Six representative statements from each category were selected for a trial instrument. The 24 items were placed on separate cards and were given to 12 educational psychologists and researchers. The directions were to divide the cards into four piles representing each of the teaching characteristics listed (Spady's four) and to record the appropriate card numbers for each teacher characteristic. No suggestion was made concerning the number of cards which might be placed in each stack. The number varied from 3 to 10. Although categories were not clearly defined, overall agreement was 77%, with individual category agreement from 72% to 80%. Five out of six items in each category had 80% or better. The items with low agreement were reevaluated. Word changes were made based on the comments of the experts during the logical validation process.

The 24 positive statements were assembled into a Q-Sort pack. Directions for administering it were developed and a form for the collection of additional data was constructed.

Pilot Test I

Students in four classes in a College of Education of a southeastern university completed the Q-Sort by ranking the cards

from 1 to 24 as to their concept of the ideal instructor for their course. The card ranked first was the most important characteristic, the last (24th) was least important. Directions involved making three stacks, one most important, one least and one in the middle, ranking within each stack and then recording numbers found on each card on a data sheet. Seventy-five graduates and 27 undergraduates completed the exercise.

Results showed the item "has competence in and knowledge about subject" was clearly the most important with a mean rank of 5.5. Means of all four categories ranked significantly higher (p. .05) across all groups than subject matter expertise and empathy. Between courses and between levels of education results were not significantly different. Interitem correlations were low even within categories indicating independence of items; at the same time correlations between each item and its category were significant.

Likert scales are typically used for student ratings. Halo effects are common. Little range has been noted in students' discrimination of variables. If a student rates an instructor high on one variable, he usually rates high on all variables and vice versa. This suggests that the Likert Scale is actually measuring one variable, teacher popularity. In an effort to combat this problem the forced choice required by ranking on a Q-sort was selected. In addition, the procedures of administering the items prior to instruction to get ideal instructor data and post course to get evaluation of what is representative of the actual instructor, and

comparing the two was indicated.

Sample size was small and no conclusions were drawn. The pilot illustrated that the evaluation was feasible to administer in about 20 minutes and that further study should follow.

Pilot Study II

The purposes of the second study were to standardize procedures for precourse and postcourse administration to obtain a sample large enough to explore construct validity, reliability and factor analysis. At the beginning of their courses, 594 graduate and undergraduate students were asked to rank the cards from most important to least important for their ideal instructor for the particular course in which they were enrolled. The instructor was asked to rank the statements as to their importance to the course. At the end of the semester the students were asked to rank the statements from most representative to least representative of the instructor for their particular course. Pre and post records were collected by a numbering system which assured that each student's responses would remain anonymous but that the pre and post for each student could be identified.

Procedures for comparing the pre (expectancy) and the post (reality) were established, also for comparing students' evaluation of what was (post) with what the instructor intended. Programs for acquiring mean ranks for items and categories on each administration, reliability (test-retest), interitem correlations and factor analysis were computed.

Some resistance to the number of cards had been noted in both pilots. Originally, the logical validation procedure had shown one weak item (which had been rewritten) in each category. These two concerns resulted in the decision to select the four items in each category which correlated highest with their category to make up the final instrument, a 16-item Q-sort entitled Student Expectancy Evaluation (SEE)(See Appendix A). As a further check the results of the factor analysis, although a questionable procedure with non independent scores-ranks were computed with the 16 remaining items. Four factors emerged with most sub-scale items loading on their respective factors. Test-retest data collected on a sample two weeks after pre-test and using the same directions yielded a reliability coefficient of .81.

Another approach to the forced choice method was tried with triads representing various combinations of those ranked items highest in each category with those ranked highest in the other categories, those lowest with the lowest, etc. A total of 72 triads made up the instrument. This was administered to 465 students on a pre-course basis. Students complained about the repetitiveness and boredom; they questioned whether their own answers were what they really meant. This approach was abandoned.

Discussion

The SEE is based on a theoretical framework suggested by Spady (1974). Logical validation was obtained by using experts to categorize statements. Empirical validation was obtained through correlational matrixes and factor analysis. Test-retest reliability (.81) is respectable for an affective instrument. Since all state-

ments are positive and the student is forced to rank all "good" characteristics, halo effects, response set and social desirability are avoided. Perhaps most important, the results consider student and teacher expectancies and give evidence of what congruence exists between what each expects and what they perceive actually occurred. Because expectancies are obtained at the beginning of the course, the instructor can modify the course early in the semester if he or she perceives it will not meet students' needs and those are valid considerations. Thus, the instrument allows for control of the expectancy variable, use of it in modifications of a course and in comparison of matches between teacher's plans and student expectancies with what students perceive they have experienced. Rankings are obtained for each course and section. Therefore, differences in curriculum, level of education and method of instruction can be reflected in the results. Some generalizations can also be made. SEE meets the criteria for a good measure in the affective domain and initial results suggests its use may invalidate current ratings scales because of their failure to account for the student expectancy variable.

A current study with 2954 subjects should confirm or disconfirm this hypothesis. Comparisons with course satisfaction are part of the analyses. Should transformed correlation coefficients between pre and post compare favorably with course satisfaction, evidence for rethinking what is being measured in instructor rating instruments will be furnished.

Evaluation Items

1. Communicates Ideas Clearly
2. Classes are Organized to Allow for Meaningful Interactions
3. Evaluates Consistently and in an Unbiased Manner
4. Presents a Well Organized Course
5. Answers Impromptu Questions Asked
6. Has Competence in, and Knowledge About, Subject
7. Identifies Basic "Truths" of Subject Area
8. Relates Knowledge of Subject Matter to Solution of Practical Problems
9. Is Interested in Whether Each Student Understands the Material
10. Is Someone with Whom a Student Can Identify and Relate
11. Appreciates Each Student's Efforts
12. Is Sensitive to the Personal Needs of Each Student
13. Stimulates Students Intellectually
14. Excites Students to Think for Themselves about Problems and Issues
15. Is Enthusiastic about the Subject and About Teaching
16. Motivates Students to Do Their Best

References

Bassin, W. A note on the biases in students' evaluations of instructors. Journal of Experimental Education, 1974, 43, 16-17.

Bausell, R. B. and Magoon, J. Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educational and Psychological Measurement, 1972, 32, 1013-1023.

Blum, M. L. An investigation of the relation existing between students' grades and their ratings of the instructor's ability to teach. Journal of Educational Psychology, 1936, 27, 217-221.

Centra, J. A. Student ratings of instruction and their relationship to student learning. American Educational Research Journal, Winter, 1977, 17-24.

Cohen, S. H. and Berger, W. G. Dimensions of students' ratings of college instructors underlying subsequent achievement on course examinations. Proceedings, 78th Annual Convention, APA, 1970, 605-606.

Elliott, O. N. Characteristics and relationships of various criteria of teaching. Unpublished doctoral dissertation, Purdue University, 1949.

Finkbeiner, C. T. Lathrop, J. S., and Scheurger, J. M. Course and instructor evaluation: some dimensions of a questionnaire. Journal of Educational Psychology, 1973, 64(2), 159-163.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. Science, Oct. 1973, 182,83.

Gage, N. The appraisal of college teaching: An analysis of ends and means. Journal of Higher Education, 1961, 32, 17-22.

Garverick, C. M. and Carter, H. D. Instructor ratings and expected grades. California Journal of Educational Research, 1962, 13, 218-221.

Herron, R. News. Central Michigan University, Mt. Pleasant, February 3, 1976.

Jaeger, R. M. and Freijo, T. D. Some psychometric questions in the evaluation of professors. Journal of Educational Psychology, 1974, 66(3), 416-423.

McDaniel, E. D. and Feldhusen, J. F. Relationships between faculty ratings and indexes of service and scholarship. Proceedings, 78th Annual Convention, APA, 1970, 619-620.

McKeachie, W. J. Student ratings of faculty. American Association of University Professors Bulletin, 1969, 55, 439-444.

McKeachie, W. J. and Solomon, D. Student ratings of instructors: a validity study. Journal of Educational Research, 1958, 51, 379-382.

Mazer, G. E. Evaluating the evaluations: a factor analysis of student ratings. Counselor Education and Supervision, Sept. 1977, 6-11.

Meredith, G. M. Dimensions of faculty-course evaluation. Journal of Psychology, 1969, 73, 27-32.

MSCellany, Sept. 1977, Measurement Services Center, Univ. of Minnesota, Minneapolis, Minnesota.

Parrish, T. S., Perrin, D. W., Prawat, R. S., and Palazzo, R. F. Student related characteristics and evaluations of instructors. Journal of Instructional Psychology, Winter, 1977, 22-27.

Perry, R. and Baumann, R. Criteria for evaluation of college teaching: Their reliability and validity at the University of Toledo. In Sockloff, A. (Ed.) Proceedings of the first invitational conference on faculty effectiveness as evaluated by students. Temple University, Philadelphia, 1973.

Rodin, M. and Rodin, B. Student evaluation of teachers. Science, Sept., 1972, 177, 1164.

Shingles, R. D. Faculty ratings: Procedures for interpreting student evaluations. American Educational Research Journal, Fall, 1977, 14:4, 459-470.

Spady, W. G. The authority system of the school and student unrest: a theoretical exploration, NSSE Yearbook on Education, 1974.

Treffinger, D. J. and Feldhusen, J. F. Predicting students' rating of instruction. Proceedings, 78th Annual Convention, APA, 1970, 621-622.

Ware, J. E., Jr. and Williams, R. G. Discriminant analysis and classification of teaching effectiveness using student rating: The search for Dr. Fox. The Rand Paper Series, Rand Corporation, Santa Monica, California, 1976.

Winne, P. H. Aptitude - treatment interactions in an experiment on teacher effectiveness. American Educational Research Journal, Fall, 1977, 14:4, 389-409.